# PROSODIC STRESS REVISITED:
# REASSESSING THE ROLE OF FUNDAMENTAL FREQUENCY

*Rosaria Silipo[1] and Steven Greenberg[2]*

[1]Nuance Communications, Inc.
1380 Willow Road, Menlo Park, CA 94025

[2] International Computer Science Institute
1947 Center Street, Berkeley, CA 94704

## ABSTRACT

In a previous study [14] we had concluded that amplitude and duration are the most important acoustic parameters underlying the patterning of prosodic stress in casually spoken American English, and that fundamental frequency ($f_o$) plays a only minor role in the assignment of stress. The current study re-examines this conclusion (using both the range and average level of $f_o$) in order to ascertain whether there may be circumstances in which fundamental frequency figures importantly in prosodic stress. Although the present results indicate that $f_o$-range is slightly more effective than average-$f_o$-level, this finding is most likely a consequence of duration-related information intrinsic to $f_o$-range, and is thus consistent with fundamental frequency playing a relatively minor role in stress assignment in naturally spoken American English.

## 1. INTRODUCTION

Prosodic stress is an integral component of spoken language, particularly for languages, such as English, that so heavily depend on it for lexical, syntactic and semantic disambiguation [10, 11]. Prosody also provides important information about the focus of the speaker's attention, highlighting for the listener what is "new" and "important" information, thus serving to facilitate processing via parsing the utterance into delimited "chunks" for reliable understanding.

Such stress-related information is derived from a complex constellation of acoustic cues associated with the duration, amplitude, and fundamental frequency ($f_o$) of syllabic sequences within an utterance [1, 5, 9, 10]. Traditionally, $f_o$ (and its perceptual correlate, pitch) has been thought to serve as the primary cue for stress in English:

> "Pitch is widely regarded, at least in English, as the most salient determinant of prominence. In other words, when a syllable or word is perceived as 'stressed' or 'emphasized,' it is pitch height or a change in pitch, more than length or loudness that is likely to be mainly responsible (see, for example, Fry 1958, Grimson 1980, pp. 222-226, Lehiste 1976, Fudge, 1984, ch. 1) ... although it is clear that stressed syllables often have greater overall acoustic intensity than weakly stressed ones, loudness seems to be the least salient and least consistent of the three parameters of pitch, duration and loudness - at least for purposes such as signaling stress." [2, p. 280]

However, it is unclear whether such statements truly apply to spontaneous speech (as opposed to scripted and non-meaningful material). In a recent study [14] we found that duration and amplitude appear to play a far more important role than $f_o$ in accounting for the stress patterns observed in spontaneous American English. The present investigation re-examines the conclusion of this earlier study in order to insure that important information associated with $f_o$ had not been unfairly disregarded.

## 2. CORPUS MATERIALS

The material used in both the previous and current studies is derived from the OGI Stories-TS corpus [3]. This corpus contains 60-second telephone monologues from several hundred different speakers covering a wide range of topics. The material was phonetically labeled and segmented at the Oregon Graduate Institute [*3*] and labeled at the prosodic level by two linguistically trained individuals at the International Computer Science Institute (cf. [15]). Three levels of stress were marked - (1) fully stressed, (2) entirely unstressed and (3) an intermediate level of stress. A total of 135 monologues (88 male and 47 female speakers) were labeled in this fashion. The agreement between the two transcribers (computed from material derived from ten separate monologues) was comparatively high, ranging from 84% for stressed syllables to 88.5% for those that are unstressed. There was considerably less agreement on the intermediate level of stress (62%), suggesting that listeners may not be able to reliably distinguish more than two distinct levels of stress in casually spoken English. The prosodically labeled material is available at http://www.icsi.berkeley.edu/~steveng/prosody.

## 3. ACOUSTIC ANALYSIS

The present analysis was performed in conjunction with the development of automatic methods for prosodically labeling spontaneous materials in American English (cf. [15]) and focuses on the relative contribution of amplitude, duration and $f_o$ to the labeling pattern of the prosodic transcribers (Figure 1). A preliminary description of this work is contained in [14]).

Information pertaining to each of the acoustic variables (amplitude, duration and $f_o$) was computationally extracted from the portion of the acoustic signal associated with the syllabic (usually vocalic) nucleus (the segmental boundaries of which had previously been delineated by transcribers at the Oregon Graduate Institute).

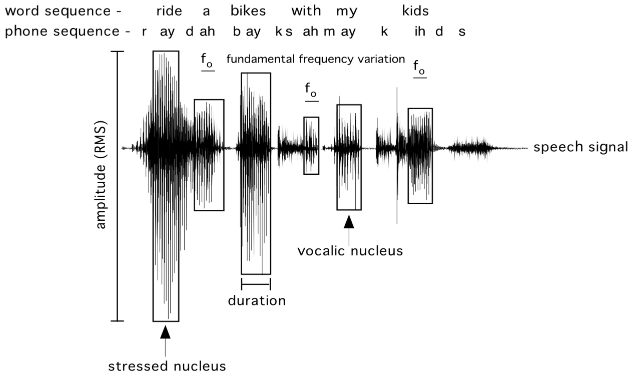

word sequence - ride a bikes with my kids
phone sequence - r ay d ah b ay k s ah m ay k ih d s
$f_o$
fundamental frequency variation
$f_o$
$f_o$
amplitude (RMS)
speech signal
vocalic nucleus
duration
stressed nucleus

**Figure 1:** A sample utterance from the OGI Stories-TS corpus illustrates the acoustic parameters (amplitude, duration and $f_o$) associated with prosodic stress.

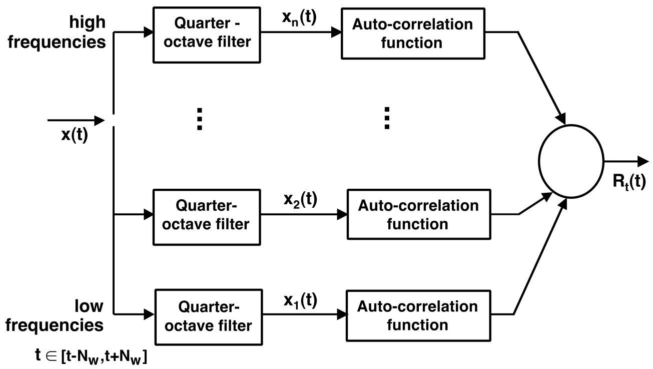


**Figure 2:** Schematic illustration of the ensemble autocorrelation analysis used to estimate $f_o$ for each frame of the syllabic nucleus (cf. [15] for details).

The amplitude was computed as the root-mean-square (rms) energy over the time interval associated with the nucleus, and the duration was defined in units of milliseconds, from the beginning of the nucleus to its end (cf. [15] for additional details).

The fundamental frequency was computed (via an ensemble autocorrelation) for successive 15 (or 25) ms windows using a frame rate of 10 (20) ms over the duration of the nucleus. The signal's spectrum was partitioned into 1/4-octave channels and the autocorrelation (cf. [8]) computed for each band every 10 (20) ms. The ensemble autocorrelation was computed by linearly summing across the individual autocorrelation functions to ascertain the peak associated with the full-bandwidth spectrum (Figure 2). A moving-average filter was applied to the ensemble autocorrelation to smooth out any singularities in the $f_o$ estimation and a baseline slope was computed via a linear fit to the $f_o$ estimates over five successive frames. The peak in the ensemble autocorrelation generally lies between 3 and 6 ms for females and between 6 and 12 ms for males.

# 4. A PROSODIC STRESS MODEL

A heuristic model for automatically labeling prosodic stress was developed as a means of combining information derived from the duration, amplitude and fundamental frequency (both average $f_o$ and $f_o$ range) of syllabic nuclei (Figure 3).

## 4.1. General Description

Each acoustic parameter was transformed into variance units and combined to form an evidence variable, $EV_k$ (Figure 4). A threshold, $T_0$, was specified for each utterance based on the proportion, *P*, of syllables that are fully stressed. This threshold was adapted to operate on a time window spanning a specified number of preceding syllables, *n* (typically, *n* is 15). This adaptive

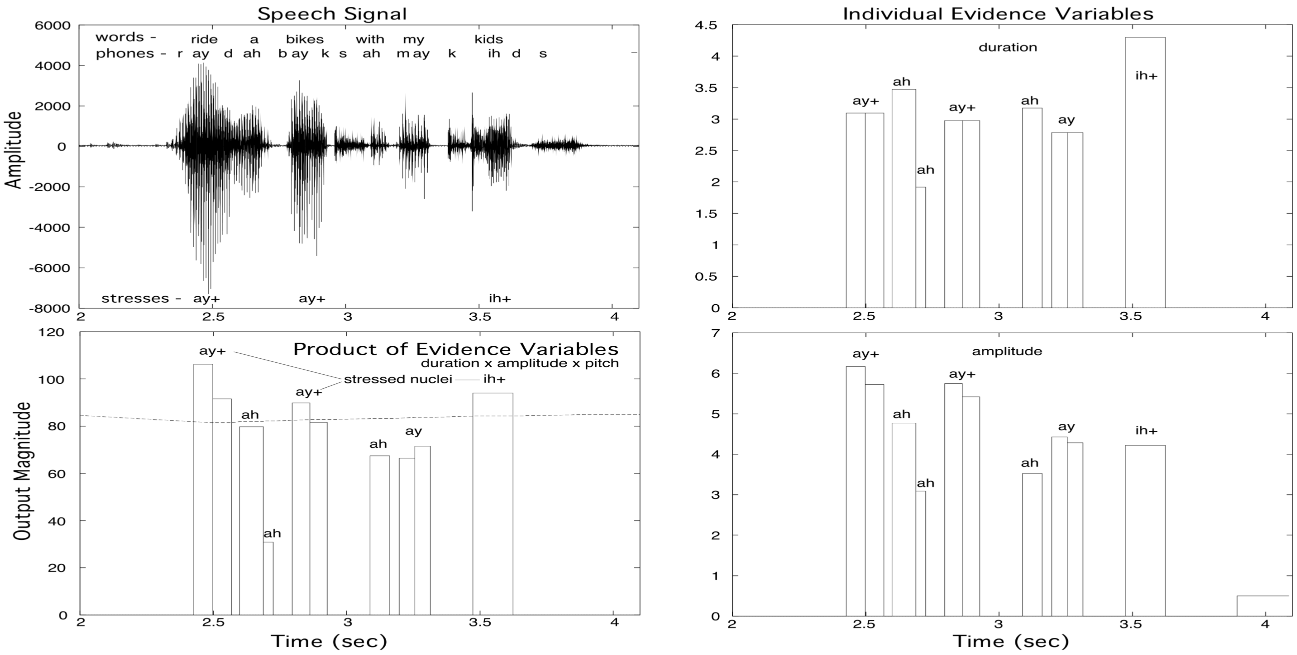


**Figure 3:** Application of the heuristic model described in Section 4 to prosodic classification of syllabic nuclei for two separate acoustic cues (amplitude and duration) as well as in combination with $f_o$.

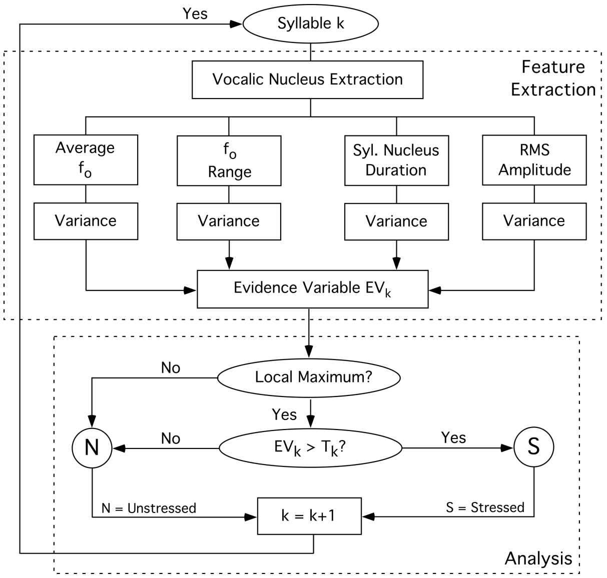


**Figure 4:** Schematic illustration of the heuristic analysis designed to ascertain the identity of acoustic cues most highly correlated with prosodic stress through the computation of an evidence variable for each parameter. See [15] for details.

threshold, $T_k$, was defined over an approximately 3-second window (but referenced to the entire 60-second utterance) in order to account for intrinsic fluctuations in speech energy over time. Only evidence variables exceeding this dynamic threshold, $T_k$, were marked as prosodically stressed. In order to ascertain whether the current value of the evidence variable, $EV_k$, represents a local maximum (i.e., a stressed syllable), it was compared with an $\alpha$ proportion of the evidence variable for the previous nucleus, $EV_{k-1}$, and a $\beta$ proportion of the evidence variable for the following vocalic nucleus, $EV_{k+1}$. $EV_k$ was marked as a local maximum if $EV_k \geq \alpha\, EV_{k-1}$ and $EV_k \geq \beta\, EV_{k+1}$. A detailed description of the model is contained in [15].

## 4.2. Training the Model

A training phase was undertaken to ascertain the optimum range of values for specific model parameters. These parameters include the percentage, *P*, of stressed syllables in an utterance, the number, *n*, of previous nuclei to use as the reference context for the acoustic features, the proportions, $\alpha$ and $\beta$, of the previous and following nuclei used to compute the local maximum, and the coefficients, *a* and *b*, for dynamic updating of the evidence variable threshold ($T_k$). Training was performed separately on two-thirds of each transcriber's data and optimum values for the model parameters determined using Receiver Operator Characteristic (ROC) curves [7].

ROC curves distinguishing stressed from unstressed syllables can be derived for different values of *n*, α, β, *a* and *b*. The proportion of stressed syllables, correctly labeled as such, is marked on the *x*-axis and the proportion of unstressed syllables correctly marked is indicated on the y-axis. The resulting ROC curve (cf. Figure 5 for an example) provides a measure of the algorithm's performance in distinguishing stressed from unstressed syllables.

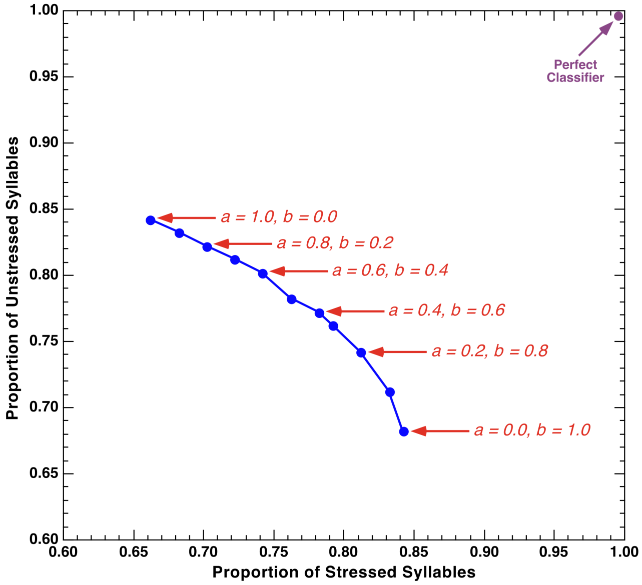


**Figure 5:** An example how a receiver operating characteristic (ROC) curve is computed when the variables, *a* and *b* (see text for detail), are systematically varied for the acoustic-parameter combination *duration, amplitude* and *f$_o$ (10 ms)*.

## 4.3. Testing the Model

A testing phase was undertaken for each evidence variable using the data (a third of the total) withheld from training. A jack-knifing procedure (rotating through different data subsets to train and test) was used in order to insure that the results are truly representative of the entire data set

In general, the optimum value for the percentage of stressed syllables, *P*, was 25%. The optimum number, *n*, of preceding syllables to include in the model's context was 15. $\alpha$ and $\beta$ were both fixed at 0.6. Parameter *a* quantifies the contribution of the initial threshold, $T_0$, while parameter *b* specifies the contribution of evidence variable over the previous 15 syllabic nuclei to the specification of the adaptive threshold, $T_k$. The specific values of *a* and *b* vary, depending on the identity of the evidence variable (cf. [15, Table 10]).

## 4.4. Evaluating the Model

The following acoustic parameters (and combinations) were evaluated using the stress-labeling model described above: (1) duration, (2) amplitude, (3) average f$_o$ using a 20-ms frame rate, (4) average f$_o$ using a 10-ms frame rate, (5) f$_o$ range (over the syllable nucleus), (6) f$_o$ range divided by duration (i.e., f$_o$ rate or duration-normalized range), (7) f$_o$ range x f$_o$ (10-ms frame rate), (8) duration x amplitude, (9) duration x f$_o$ range, (10) duration x average f$_o$ (10 ms), (11) average f$_o$ (10 ms) x amplitude, (12) f$_o$ range x amplitude, (13) f$_o$ range x amplitude x duration, (14) average f$_o$ (10 ms) x amplitude x duration and (15) f$_o$ range x average f$_o$ (10 ms) x amplitude x duration.

The results of the ROC analyses are illustrated in Figures 6 - 11. Figures 6 and 7 show the algorithm's performance when only a single acoustic parameter is used for labeling, while Figures 8-11 pertain to using these same acoustic attributes in combination with each other. The even-numbered figures apply to the labeling data of

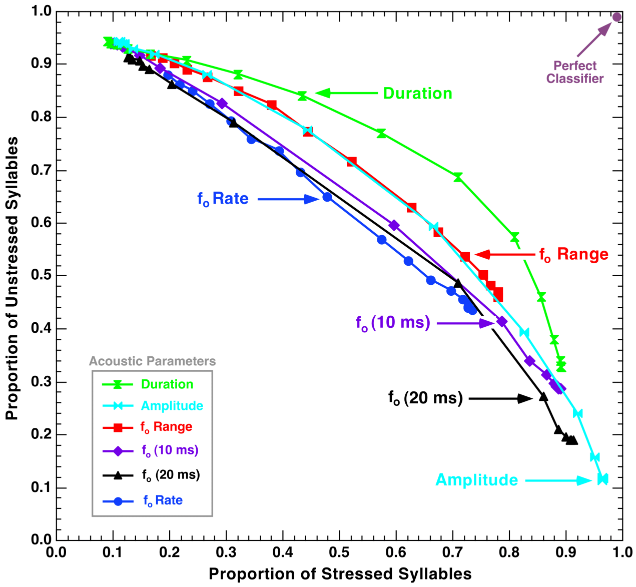


**Figure 6:** An ROC curve delineating the performance of acoustic parameters used in isolation and derived from the stress-labeling data of Transcriber 1. Duration of the syllabic nucleus is the single most effective parameter. See Table 1 for details.

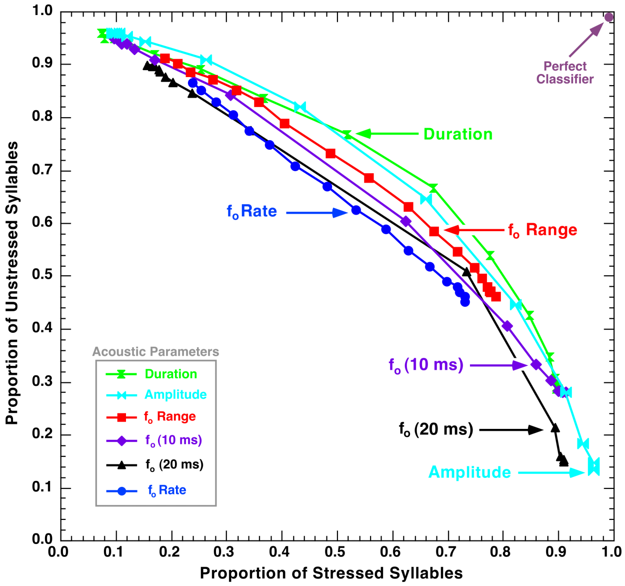


**Figure 7:** An ROC curve delineating the performance of acoustic parameters used in isolation and derived from the stress-labeling data of Transcriber 2. Note the similarity of the parameters' rank order to those of Transcriber 1.

Transcriber 1, while the odd-numbered counterparts pertain to the labels of Transcriber 2.

Duration of the syllabic nucleus is the single most important parameter for predicting the labeling pattern of both transcribers, while the parameters associated with $f_o$ are generally far less effective. The primary difference in the labeling pattern of the two transcribers involves the relative contribution of fundamental frequency variation within the syllabic nucleus. For Transcriber 1 this parameter is tied for second place (with amplitude), while for Transcriber 2 its performance falls well behind that of both duration and amplitude. For both transcribers, the other $f_o$-based parameters appear to play a tertiary role when used as the sole basis of the labeling algorithm, and the rank order of effectiveness for the acoustic parameters is similar for both sets of data (save for the virtual tie between $f_o$-range and amplitude in Transcriber 1's data).

The effectiveness of $f_o$-range relative to the other $f_o$-based parameters may be a consequence of a "hidden" role played by duration. Long syllabic nuclei provide a potentially greater opportunity for variation in $f_o$ than their briefer counterparts. If $f_o$ range is normalized by duration (for each syllabic nucleus) the gain in performance is entirely wiped out (cf. $f_o$-Rate in Figures 6 and 7), suggesting that $f_o$-related information is truly of secondary (if not tertiary) importance with respect to the stress-labeling pattern associated with the OGI Stories corpus. The data in Table 1 (detailing the optimum performance associated with each parameter) are consistent with this conclusion.

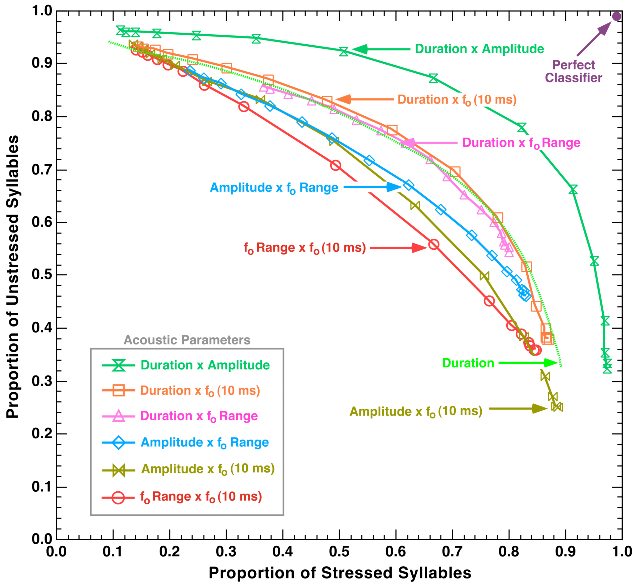
A defining characteristic of prosodic stress is its multifaceted nature. Stress is rarely based solely on a single acoustic attribute, and therefore it is of interest to ascertain how well the heuristic model can perform when two (Figures 8 and 9) or more (Figures 10 and 11) acoustic characteristics are combined. When two parameters are combined, the best performance is achieved by the product of *duration* and *amplitude*. There is no other two-parameter combination that comes even close to achieving its level of performance. It is also of interest that duration, *by itself* (indicated on the figures by a dotted green curve), provides as good a basis for labeling stress as any two-parameter combination (other than when it is itself paired with amplitude). None of the $f_o$-based parameters gains significantly in performance when paired with another parameter, lending further support to the conclusion that fundamental frequency is, at best, a secondary attribute with respect to stress (at least for the OGI Stories corpus).

Figures 10 and 11 illustrate the impact on predicting stress labels when three or more parameters are combined. No three- or four-parameter combination equals the performance of the product of *duration* and *amplitude* (marked by the dashed spearmint-colored curve on Figures 10 and 11). When $f_o$-based parameters are combined with amplitude and duration the result is a *decline* in performance, particularly when $f_o$ range is involved. This pattern reinforces the conclusion that $f_o$ plays a relatively minor role in the stress assignment of casually spoken American English and favors a model in which amplitude and duration play a dominant role.

## 5. STATISTICAL DECISION TREES

Another means by which to assess the relative contributions of duration, amplitude and $f_o$ is through the application of statistical decision trees [12, 13]. This technique partitions the material in such a fashion as to maximize the entropy associated with each "cut" through the data. The goal of such an analysis is to derive a set of features (in this instance, acoustic parameters) that account as completely as possible for the stress-labeling pattern of the two transcribers.

Figure 12 illustrates the patterning of acoustic parameters associated with the stress-labeling data of Transcriber 1. Duration and amplitude dominate the tree for all but the lowest nodes. The



**igure 8:** An ROC curve delineating the performance of acoustic arameters *pairs* derived from stress-labeling data of Transcriber 1. he product of *duration* and *amplitude* yields the best performance.

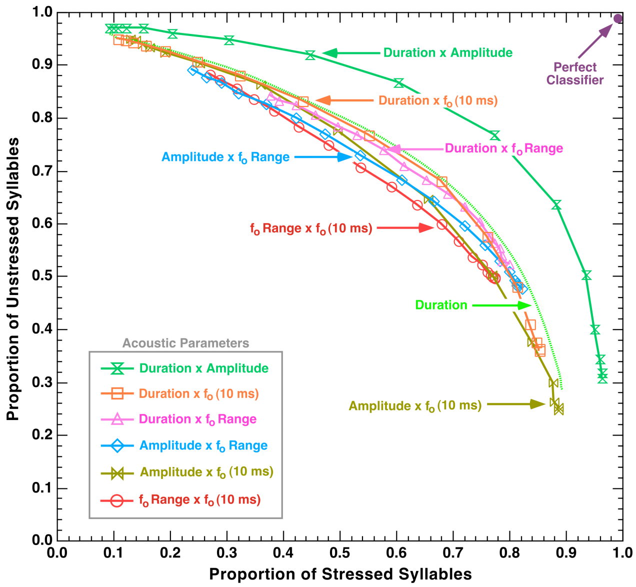


**Figure 9:** An ROC curve delineating the performance of acoustic parameters *pairs* derived from stress-labeling data of Transcriber 2. The rank order of the parameters is similar to that of Transcriber 1.

decision-tree pattern of Transcriber 2 is similar except for a proliferation of $f_o$-dominated nodes at the lowest tiers of the tree.

The decision trees are capable of classifying the stress patterns associated with the OGI Stories corpus at a level comparable to that of the ROC-based heuristic model (cf. [15]).

## 6. CONCLUSIONS

Both the heuristic-model and decision-tree analyses strongly imply that fundamental frequency plays a relatively minor role in the assignment of prosodic stress in casually spoken American English, and that amplitude and duration are the primary acoustic parameters associated with the patterning of stress-relevant cues in spontaneous material such as the OGI Stories corpus (and is consistent with the conclusions of [9] for a comparable corpus of spoken Dutch). It is of particular interest that amplitude and duration are the only two parameters that significantly enhance stress-labeling performance when used in combination. And it is also of significance that the only $f_o$-related parameter providing superior labeling performance when used in isolation ($f_o$ range) so heavily depends on duration.

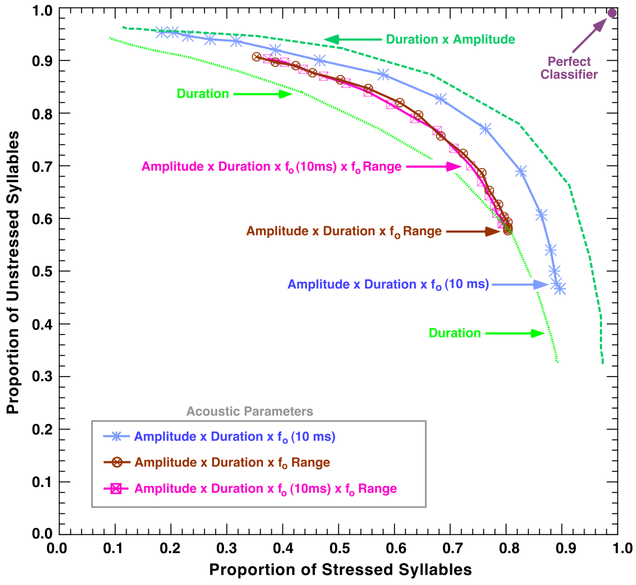


**Figure 10:** An ROC curve delineating the performance of acoustic parameters *combinations* (three or four) derived from stress-labeling data of Transcriber 1. No single combination is as effective as the product of *amplitude* and *duration*.

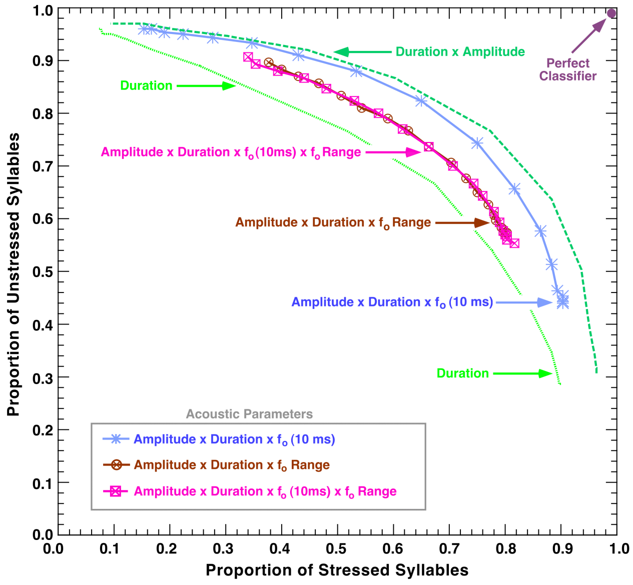


**Figure 11:** An ROC curve delineating the performance of acoustic parameters *combinations* (three or four) derived from stress-labeling data of Transcriber 2. Note the similarity of the curves with those shown in Figure 10 for Transcriber 1.

| Acoustic Parameter | | | | Percent Concordance | | |
|---|---|---|---|---|---|---|
| Duration | Amplitude | $f_0$ (Mean) | $f_0$ (Range) | Stressed | Unstressed | Neither |
| X | | | | 69.0 | 68.0 | 58.0 |
| | X | | | 63.5 | 65.0 | 49.0 |
| | | X | | 61.5 | 60.0 | 49.5 |
| | | | X | 62.5 | 63.0 | 51.5 |
| X | X | | | 79.0 | 78.0 | 60.0 |
| X | | X | | 69.0 | 69.0 | 57.5 |
| X | | | X | 69.5 | 66.0 | 59.0 |
| | X | X | | 64.5 | 64.0 | 49.5 |
| | X | | X | 68.5 | 62.5 | 55.5 |
| | | X | X | 63.5 | 65.5 | 52.5 |
| X | X | X | | 75.5 | 75.0 | 56.0 |
| X | X | | X | 71.0 | 72.5 | 57.0 |
| X | X | X | X | 70.5 | 70.5 | 55.0 |
| $f_0$ (Mean - 20-ms frame rate) | | | | 70.0 | 54.0 | 51.5 |
| $f_0$-Range / Duration (normalized) | | | | 58.5 | 58.0 | 51.5 |

**Table 1:** The heuristic model's average concordance with Transcribers 1 and 2 for stress-labeling the OGI Stories corpus on the basis of fifteen different acoustic parameters (and combinations). The concordance shown applies to data shown in Figures 6-11. The most effective acoustic parameters are shown in green (single attributes), blue (parameter pairs) and magenta (parameter combinations).

## ACKNOWLEDGEMENTS

The authors thank Jeff Good and Joy Hollenback for stress-labeling the OGI Stories material. This study was supported by the U.S. Department of Defense.

## REFERENCES

1. Beckman, M., *Stress and Non-Stress Accent.* Fortis, Dordrecht, 1986.

2. Clark, J. and Yallup, C., *Introduction to Phonology and Phonetics*, Blackwell, Oxford, 1990.

3. Cole, R., Fanty, M., Noel, M. and Lander, T., "Telephone Speech Corpus Development at CSLU," *Proc. Int. Conf. Spoken Lang. Proc.*, 1994.

4. Fry, D., "Experiments in the perception of stress," *Lang. Speech*, 1, 126-152,

5. Fudge, E., *English Word-Stress,* Allen and Unwin, London, 1984.

6. Gimson, A., *An Introduction to the Pronunciation of English (3rd ed.),* Edward Arnold, London, 1980.

7. Green, D. and Swets, J., *Signal Detection Theory and Psychophysics*, Wiley, New York, 1966.Lehiste, I., Suprasegmentals, MIT Press, Cambridge, 1970. on page 6

8. Hess, W., *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer Verlag, Berlin, 1983.

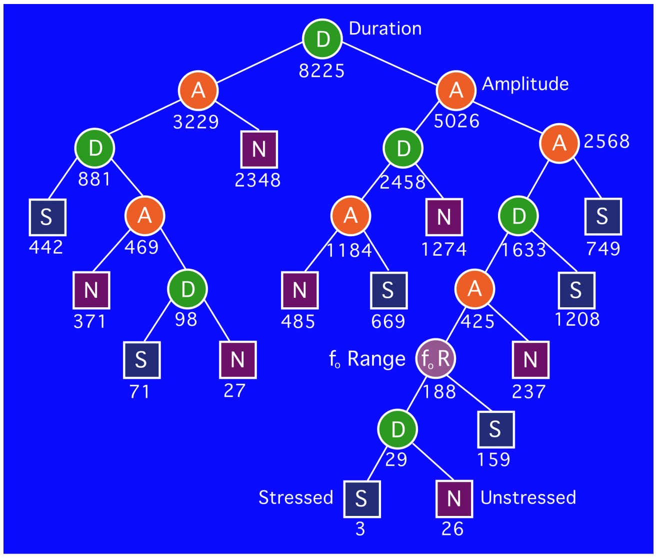


**Figure 12:** A statistical decision tree generated to account for a portion of the stress-labeling data produced by Transcriber 1. The tree's nodes are marked by circles, color-coded by acoustic parameter. The leaves of the tree are marked by squares. Associated with each leaf and node is the number of labels accounted for by a particular branch of the tree. Duration and amplitude dominate the tree at all but the lower nodes. See [15] for further details.

9. Kuijk, D. van and Boves, L., "Acoustic characteristics of lexical prominence in continuous telephone speech," *Speech Communication*, 27, 95-111, 1999.

10. Lehiste, I., *Suprasegmentals*, MIT Press, Cambridge, 1970.

11. Lehiste, I, "Suprasegmentals," in N. Lass (ed.) *Contemporary Issues in Experimental Phonetics*, Academic Press, New York.

12. Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.

13. Quinlan, J.R., "Induction of Decision Trees," *Machine Learning*, pp. 81-106, 1986.

14. Silipo, R. and Greenberg, S., "Automatic transcription of prosodic prominence for spontaneous English discourse," *Proc. XIVth Int. Cong. Phon. Sci.*, pp. 2351-2354, 1999 (available from http://www.icsi.berkeley.edu/~steveng).

15. Silipo, R. and Greenberg, S., *Automatic Detection of Prosodic Stress in American English Discourse*, Technical Report TR-00-001 (29 pages), International Computer Science Institute, Berkeley, 2000 (available from http://www.icsi.berkeley.edu/techreports/2000.html)